INFERENCE TO THE BEST EXPLANATION MADE INCOHERENT*

Proponents of Inference to the Best Explanation (IBE) claim that our inferences should give explanatory considerations a central role. Beyond this general agreement, however, they have differed on precisely *how* explanation should inform inference. A particular area of controversy has been the relation of IBE to Bayesianism. Should IBE be formulated in terms of full beliefs, as in traditional epistemology, or in terms of degrees of belief, as in Bayesian epistemology? If it is formulated in the latter way, is it compatible with Bayesian epistemology?

In this essay, I advance a new argument against non-Bayesian formulations of IBE, which include both traditional formulations of IBE in terms of full belief and non-Bayesian formulations of IBE in terms of degrees of belief. I show that in some instances, IBE for full belief licenses deductively inconsistent inferences from the same evidence. In similar instances, following non-Bayesian IBE updating rules for degrees of belief leads to probabilistically incoherent credences.

In section I, I present the problem for traditional formulations of IBE in terms of full belief. In section II, I present the problem for non-Bayesian formulations of IBE in terms of degrees of belief. In section III, I consider some possible responses on behalf of the proponent of these formulations. Finally, in section IV, I conclude with some reflections on what a *Bayesian* form of IBE could look like.

I. IBE FOR FULL BELIEF

Traditionally, proponents of IBE have formulated it in terms of full belief or in terms of similar binary notions, such as acceptance. The most common formulation of IBE is that suggested by the phrase 'inference to the best explanation': it is a procedure on which, if H is the best explanation of one's evidence, then one infers that H is true. This formulation is endorsed by the following authors:

> In making this inference [that is, in inferring to the best explanation] one infers, from the fact that a certain hypothesis would explain the evidence, to the truth of that hypothesis. In general, there will be several

*I am grateful to Igor Douven, Daniel Immerman, Jonah Schupbach, and an anonymous reviewer for this JOURNAL for extremely helpful comments on earlier drafts. I would also like to thank audience members at Notre Dame and Universidad de Los Andes, especially Lane DesAutels and Mark Satta, for helpful feedback on presentations of this project.

0022-362X/17/1405/251–273

> hypotheses which might explain the evidence, so one must be able to reject all such alternative hypotheses before one is warranted in making the inference. Thus one infers, from the premise that a given hypothesis would provide a "better" explanation for the evidence than would any other hypothesis, to the conclusion that the given hypothesis is true.[1]
>
> Inference to the best explanation is the procedure of choosing the hypothesis or theory that best explains the available data.[2]
>
> IBE authorises the acceptance of a hypothesis H, on the basis that it is the best explanation of the evidence.[3]
>
> IBE can be written as follows: It is reasonable to believe that the best available explanation of any fact is true.[4]
>
> [In inference to the best explanation,] one concludes that something is the case on the grounds that this best explains something else one believes to be the case.[5]

It is obvious that this form of IBE can lead to inconsistent conclusions when applied multiple times to different items of evidence. For example, one hypothesis may best explain evidence E_1, whereas an inconsistent hypothesis may best explain additional evidence E_2. One might have thought, however, that IBE cannot lead to inconsistent conclusions when applied to the *same* evidence, so that we can avoid inconsistency by only drawing inferences from our *total evidence* (for example, $E_1\&E_2$). Against this idea, I will now show that in some cases, the above form of IBE can license inconsistent inferences from the same evidence.

My argument is based on the following fact about human reasoning: we offer and search for explanations at multiple levels. We can, with equal propriety, explain the presence of a trait in a biological population by positing that its members possess a gene that has been known to lead to that trait in similar organisms, or by telling a story about how that trait helped the population's ancestors survive and reproduce. We can explain the invasion of one country by another by appealing to the beliefs and desires of the politicians of the invading

[1] Gilbert Harman, "The Inference to the Best Explanation," *Philosophical Review*, LXXIV, 1 (January 1965): 88–95, at p. 89.

[2] Jonathan Vogel, "Inference to the Best Explanation," in Edward Craig, ed., *Routledge Encyclopedia of Philosophy* (London: Routledge, 1998).

[3] Stathis Psillos, "Inference to the Best Explanation and Bayesianism," in F. Stadler, ed., *Induction and Deduction in the Sciences* (Dordrecht: Kluwer, 2004), p. 83 (italics omitted).

[4] Alan Musgrave, "Experience and Perceptual Belief," in Zuzana Parusniková and Robert S. Cohen, eds., *Rethinking Popper* (Dordrecht: Springer, 2009), p. 17 (italics omitted).

[5] John Greco, "Inference to the Best Explanation," in Robert Audi, ed., *The Cambridge Dictionary of Philosophy, Third Edition* (New York: Cambridge University Press, 2015), p. 510.

country, current geopolitical factors, or historical tensions between the countries. Some explanations we offer or consider are more distal, and some are more proximate; some are more local, and some are more global. Despite their differences, all of these types of explanations are generally recognized as legitimate, and they are all types that we might be interested in inferring from the data available to us. Although explanations at multiple levels of explanation may conflict, they need not: the above explanations of the biological trait and national invasion could all be true.[6]

However, in cases where explanations at multiple levels are *not* compatible, IBE can license inconsistent inferences from the same data. Consider a case in which I tell you the following. I have four urns with the following contents:

U_1 = 4 white balls
U_2 = 2 black balls, 2 white balls
U_3 = 3 black balls, 1 white ball
U_4 = 4 black balls

I am going to set one of these urns in front of you, and you are going to draw a ball from it. I will select the urn by flipping a coin twice. If it lands heads the first time, I will flip it again to select between U_1 and U_4. If it lands tails the first time, I will flip it again to select between U_2 and U_3.

In this case, we have the following two partitions of possibilities:[7] {U_1, U_2, U_3, U_4} and {Heads, Tails} (where the latter corresponds to the outcome of the first coin flip). We also have the following material equivalencies:

Heads = $U_1 \vee U_4$
Tails = $U_2 \vee U_3$

Now suppose that I flip the coin twice, and I set the chosen urn in front of you. I do not tell you which urn is in front of you or the outcome of the coin flip. Your job is to infer that information by sampling from the urn. Here there are two possibilities: {Black, White}. You draw out a black ball. What should you conclude?

According to the simple version of IBE, you should infer the best explanation of Black. What is that? If we are concerned with the question of which urn is in front of you, then U_4 is the best explanation, for

[6] Peter Lipton makes a similar point in his *Inference to the Best Explanation*, 2nd ed. (London: Routledge, 2004), pp. 62–63, writing that in cases like those I have described, "in spite of the suggestion of uniqueness that the word 'best' carries, Inference to the Best Explanation should be construed so as to allow multiple explanations."

[7] A partition {H_1, …, H_n} is a set of mutually exclusive and jointly exhaustive alternatives.

that urn has only black balls and U_4 does not start out as any more or less plausible than the other urn hypotheses. However, if we are concerned with the question of whether I flipped heads or tails, then Tails is the better explanation. Tails makes it more likely that you would draw black than does Heads, and neither Heads nor Tails is initially more or less plausible than the other.

More formally, these judgments about explanation follow from this principle about the conditions under which one hypothesis better explains the evidence than another:

(1) If H_1 and H_2 are both potential explanations of E, $P(H_1) = P(H_2)$, and $P(E|H_1) > P(E|H_2)$, then H_1 explains E better than H_2.

What (1) says is that if two potential explanations of the evidence start out equally credible and one makes the evidence more likely, then that one explains the evidence better.

In the case at hand,

$$P(\text{Heads}) = P(\text{Tails}),$$

and

$$P(\text{Black}|\text{Tails}) = 5/8 > P(\text{Black}|\text{Heads}) = 1/2.$$

Likewise,

$$P(U_1) = P(U_2) = P(U_3) = P(U_4) = 1/4,$$

and

$$P(\text{Black}|U_4) = 1 > P(\text{Black}|U_3) = 3/4 > P(\text{Black}|U_2) = 1/2 > P(\text{Black}|U_1) = 0.$$

Hence, by (1), Tails explains Black better than does Heads, and U_4 explains Black better than does U_3 or U_2. As such, the simplest form of IBE would direct you to infer both U_4 and Tails. However, U_4 and Tails are inconsistent. Therefore, following this form of IBE leads to deductively inconsistent beliefs in this case.

Many proponents of IBE have offered more sophisticated versions of the view as applied to full belief. For example, Theo Kuipers has defended a version of IBE on which, given a set of candidate explanations $\{H_1, \ldots, H_n\}$ of E, one infers not that the best explanation of E among those is *true* but that it is the *closest* of these hypotheses to the truth.[8]

[8] Theo A. F. Kuipers, "Approaching the Truth with the Rule of Success," *Philosophia Naturalis*, XXI (1984): 244–53; Theo A. F. Kuipers, "Naïve and Refined Truth Approximation," *Synthese*, XCIII, 3 (December 1992): 299–341. See also Igor Douven, "Abduction," in Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Spring 2011 Edition), URL = http://plato.stanford.edu/archives/spr2011/entries/abduction/, section 2.

Following this version of IBE will lead to inconsistent beliefs about the closeness to truth of the hypotheses in the present case. Exactly one of {Heads, Tails} is true, and exactly one of $\{U_1, U_2, U_3, U_4\}$ is true. Tails is closer to the truth than is Heads just in case Tails is true. If Tails is true, then one of U_2 or U_3 is closer to the truth than U_4, because one of U_2 or U_3 is, simply, true. Hence, if one's belief that Tails is closer to the truth than Heads is true, then one's belief that U_4 is closer to the truth than the other urn hypotheses is false. Conversely, if U_4 is closer to the truth than the other urn hypotheses, that is because it is, simply, true; in this case, Heads is closer to the truth than is Tails, because Heads is true and Tails is false.[9]

Other philosophers have argued that we should only infer the best explanation when that explanation is *satisfactory* or *good enough*.[10] Because of the imprecision of the notion of 'good enough', it is difficult to tell whether this version of IBE would license inference to both U_4 and Tails in the above case. Suppose that it does not, however. It is plausible that the requirement that H be a good enough explanation of E will not rule out the above kind of scenario more generally unless it always implies that, given E, H is more probable than not. (In the above scenario, $P(U_4|\text{Black}) = 4/9$.[11] So U_4 is less probable than not even after you draw black. More generally, A and B can only be inconsistent if one of them has a probability less than or equal to .5.)

Requiring that $P(H|E) > .5$, however, will not preclude inconsistent beliefs in scenarios with more than two levels of explanations. To demonstrate this, I will now describe a case with three levels of

[9] Most theories of closeness to the truth, including Kuipers's in "Naïve and Refined Truth Approximation," *op. cit.*, combine a "truth factor" with a "content factor." (For an overview of such theories, see Graham Oddie, "Truthlikeness," in Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), URL = http://plato.stanford.edu/archives/spr2016/entries/truthlikeness/.) Roughly speaking, the truth factor measures how close a proposition comes to being true, whereas the content factor measures how informative a proposition is. The idea in combining them is that we are interested in how close a proposition is to the *whole* truth, and (other things equal) informative propositions come closer to the whole truth than do uninformative propositions (such as tautologies). In the text, I focused solely on the truth factor because Heads and Tails are equally informative and the urn hypotheses are equally informative. As such, the only thing that makes a difference in how close each one is to *the* truth is how close they are to being true—that is, the truth factor.

[10] Alan Musgrave, "The Ultimate Argument for Scientific Realism," in Robert Nola, ed., *Relativism and Realism in Science* (Dordrecht: Kluwer, 1988), pp. 229–52; Peter Lipton, "VI*—Is the Best Good Enough?," *Proceedings of the Aristotelian Society*, XCIII, 1 (1993): 89–104; Lipton, *Inference to the Best Explanation*, *op. cit.*, pp. 151–63. See also Douven, "Abduction," *op. cit.*, section 2.

[11] See the application of Bayes's rule to the urn hypotheses in section II for this calculation.

explanation, in which the best explanations at each level each have posterior probability 2/3 and yet are jointly inconsistent.

A jury is deliberating about the killing of Mr. Boddy in the study with the revolver. The jury is trying to answer three questions: (a) Who is the killer? (b) What was the motive? (c) Was the killing planned? We have the following three partitions corresponding to these questions: {Plum, Green}, {Defense, Money}, and {Planned, ~Planned}. Let E be the evidence that Boddy was killed with the revolver. Suppose that (before learning E) the jury's background knowledge imposes the probability distribution on which P(Green&Money&~Planned&~E) = P(Plum&Defense&~Planned&E) = P(Plum&Money&Planned&E) = P(Green&Defense&Planned&E) = 1/4. Where each quadrant has probability 1/4, this probability distribution can be represented as follows:

Green&Money&~Planned&~E	Plum&Defense&~Planned&E
Plum&Money&Planned&E	Green&Defense&Planned&E

So,

$$P(E|Plum) = P(E|Defense) = P(E|Planned) = 1$$
$$> P(E|Green) = P(E|Money) = P(E|\sim Planned) = 1/2,$$

$$P(Plum) = P(Green) = P(Defense) = P(Money) = P(Planned) = P(\sim Planned) = 1/2,$$

$$P(Plum|E) = P(Defense|E) = P(Planned|E) = 2/3.$$

In other words, Plum (but not Green) would definitely use the revolver, the killer would definitely use the revolver in self-defense (but might use another weapon for money), and a premeditated homicide (but not a spontaneous one) would definitely be committed with the revolver. But Plum would not plan to kill someone in self-defense. (Green might, for some reason—perhaps Green had an antecedent reason to believe that Boddy was out to get him.) So in this case, that Plum did it, that it was in self-defense, and that the killing was planned are all better explanations of the killer using the revolver than are their rivals. In addition, all have posterior probabilities above .5. Nevertheless, these explanations are jointly inconsistent, and P(Plum&Defense&Planned|E) = 0.

If we raise the probabilistic threshold implied by "H is a good enough explanation of E" above 2/3, then we can rule out the explanations being good enough in the above case. But by adding in further levels of explanation, we can devise a case of the above sort for any threshold that falls short of 1. I conclude that adding to our

explanationist inference rule the requirement that H be a good enough explanation of E will not preclude the same evidence from warranting inconsistent inferences in some cases.[12]

II. IBE FOR DEGREES OF BELIEF

In a series of essays, Igor Douven, Sylvia Wenmackers, and Jonah Schupbach have explored a particular version of IBE, which I will, following Douven and Wenmackers,[13] call IBE*. IBE* is a non-Bayesian updating rule that gives "bonus" points to more explanatory hypotheses.

IBE* was first discussed, albeit in a less precise form, by Bas van Fraassen, who argued that an agent who follows IBE* will end up with diachronically incoherent credences. Van Fraassen took this to show that IBE and Bayesianism are incompatible.[14] Although not endorsing IBE* as the only reasonable explication of IBE, Douven[15] subsequently argued that this incompatibility does not refute IBE* as a legitimate updating rule, contending that explanationists can adopt IBE* and nevertheless defend themselves against van Fraassen's diachronic Dutch book argument as well as Hannes Leitgeb and Richard Pettigrew's[16] inaccuracy-minimization argument. More recently, Douven, Wenmackers, and Schupbach have argued that IBE* is more attractive than Bayes's rule in certain respects, including speed of convergence to the truth,[17] performance in a social setting,[18] and accuracy as a description of people's *actual* probabilistic updating.[19]

[12] One might argue that cases with more than two levels of explanation are less troubling inasmuch as the lottery and preface paradoxes already show that deductive inconsistency is sometimes rationally permissible for larger sets of beliefs. I consider this response in section III.4.

[13] Igor Douven and Sylvia Wenmackers, "Inference to the Best Explanation versus Bayes's Rule in a Social Setting," *British Journal for the Philosophy of Science*, LXVIII, 2 (June 2017): 535–70.

[14] Bas van Fraassen, *Laws and Symmetry* (Oxford: Oxford University Press, 1989), chapter 7, section 4.

[15] Igor Douven, "Inference to the Best Explanation Made Coherent," *Philosophy of Science*, LXVI (September 1999): S424–35; Igor Douven, "Inference to the Best Explanation, Dutch Books, and Inaccuracy Minimisation," *The Philosophical Quarterly*, LXIII, 252 (July 2013): 428–44.

[16] Hannes Leitgeb and Richard Pettigrew, "An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy," *Philosophy of Science*, LXXVII, 2 (April 2010): 236–72.

[17] Douven, "Inference to the Best Explanation, Dutch Books, and Inaccuracy Minimisation," *op. cit.*

[18] Douven and Wenmackers, "Inference to the Best Explanation versus Bayes's Rule in a Social Setting," *op. cit.*

[19] Igor Douven and Jonah N. Schupbach, "Probabilistic Alternatives to Bayesianism: The Case of Explanationism," *Frontiers in Psychology*, VI (2015); Igor Douven and Jonah N. Schupbach, "The Role of Explanatory Considerations in Updating," *Cognition*, CXLII (September 2015): 299–311.

Douven's defenses of IBE* focus on objections to IBE* that turn on its leading to diachronically incoherent credences. In this section, I show that in some cases, using IBE* to update on new evidence leads to *synchronic* as well as diachronic incoherence. IBE* is thus not only "non-Bayesian" in that it violates Bayesian conditionalization; it is non-probabilist in that it violates the requirement that an agent's credences be probabilities. In addition, we will see that other non-Bayesian formulations of IBE for degrees of beliefs, including ones suggested by Douven as alternatives to IBE*,[20] have this same consequence.

To see how IBE* works, consider updating your credences about my coin flip in the original urn case. We have the following possibilities:

U_1 = 4 white balls
U_2 = 2 black balls, 2 white balls
U_3 = 3 black balls, 1 white ball
U_4 = 4 black balls
Heads = $U_1 \vee U_4$
Tails = $U_2 \vee U_3$

You have drawn a black ball from the urn. What should your new credence that tails was flipped be?

According to Bayes's rule (also called Bayesian conditionalization),[21] where Cr(Tails) is your original credence that the coin landed tails and Cr_{new}(Tails) is your credence that it landed tails after updating,[22]

$$Cr_{new}(\text{Tails}) = Cr(\text{Tails}|\text{Black})$$

$$= \frac{Cr(\text{Tails})Cr(\text{Black} \mid \text{Tails})}{Cr(\text{Tails})Cr(\text{Black} \mid \text{Tails}) + Cr(\sim\text{Tails})Cr(\text{Black} \mid \sim\text{Tails})}$$

$$= \frac{\left(\frac{1}{2}\right)\left(\frac{5}{8}\right)}{\left(\frac{1}{2}\right)\left(\frac{5}{8}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)} = \frac{\frac{5}{16}}{\frac{5}{16} + \frac{1}{4}} = \frac{\frac{5}{16}}{\frac{9}{16}} = \frac{5}{9} \approx .56$$

More generally, where $\{H_1, \ldots, H_n\}$ is a partition of hypotheses, Bayes's rule says that

[20] Igor Douven, "Inference to the Best Explanation: What Is It? And Why Should We Care?," in Ted Poston and Kevin McCain, eds., *Best Explanations: New Essays on Inference to the Best Explanation* (New York: Oxford University Press, forthcoming), chapter 2.

[21] I follow Douven in using the term "Bayes's rule," but stress that this rule should not be confused with Bayes's theorem. The latter is a theorem of the probability calculus and is not (or should not be) controversial. It is the second equality below, which expresses P(Heads|Black) as a function of P(Heads), P(Black|Heads), and P(Black|~Heads). The former is a philosophically controversial claim that says that an agent's credence in H after getting evidence E should be equal to her former credence in H conditional on E.

[22] I switch from P(.) to Cr(.) here because I will be discussing some credence functions that are probabilistically incoherent and hence are not probability functions.

$$\mathrm{Cr}_{\mathrm{new}}(\mathrm{H}_i) = \mathrm{Cr}(\mathrm{H}_i \mid \mathrm{E}) = \frac{\mathrm{Cr}(\mathrm{H}_i)\mathrm{Cr}(\mathrm{E} \mid \mathrm{H}_i)}{\sum_j \left[\mathrm{Cr}(\mathrm{H}_j)\mathrm{Cr}(\mathrm{E} \mid \mathrm{H}_j)\right]}$$

According to IBE*, by contrast,

$$\mathrm{Cr}_{\mathrm{new}}(\mathrm{H}_i) = \frac{\mathrm{Cr}(\mathrm{H}_i)\mathrm{Cr}(\mathrm{E} \mid \mathrm{H}_i) + f(\mathrm{H}_i, \mathrm{E})}{\sum_j \left[\mathrm{Cr}(\mathrm{H}_j)\mathrm{Cr}(\mathrm{E} \mid \mathrm{H}_j) + f(\mathrm{H}_j, \mathrm{E})\right]}$$

$f(\mathrm{H}_i, \mathrm{E})$ is a function that assigns a non-negative bonus b to the hypothesis H_i that best explains the evidence, and 0 to all other hypotheses. (IBE* and Bayes's rule are equivalent just in case this bonus is 0.) If we set the bonus b to 1/8, then, given that Tails explains drawing black better than does Heads, IBE* would calculate your new credence in the present case as follows:

$$\begin{aligned}
\mathrm{Cr}_{\mathrm{new}}(\mathrm{Tails}) &= \frac{\mathrm{Cr}(\mathrm{Tails})\mathrm{Cr}(\mathrm{Black} \mid \mathrm{Tails}) + \frac{1}{8}}{\mathrm{Cr}(\mathrm{Tails})\mathrm{Cr}(\mathrm{Black} \mid \mathrm{Tails}) + \mathrm{Cr}(\mathrm{Heads})\mathrm{Cr}(\mathrm{Black} \mid \mathrm{Heads}) + \frac{1}{8}} \\
&= \frac{\frac{5}{16} + \frac{2}{16}}{\frac{9}{16} + \frac{2}{16}} \\
&= \frac{7}{11} \approx .64
\end{aligned}$$

A parallel calculation for $\mathrm{Cr}_{\mathrm{new}}(\mathrm{Heads})$ would find that it is equal to 4/11, or approximately .36.

What about the urn hypotheses? According to Bayes's rule, your new credence in U_4 is:

$$\begin{aligned}
\mathrm{Cr}_{\mathrm{new}}(\mathrm{U}_4) &= \frac{\mathrm{Cr}(\mathrm{U}_4)\mathrm{Cr}(\mathrm{Black} \mid \mathrm{U}_4)}{\sum_i [\mathrm{Cr}(\mathrm{U}_i)\mathrm{Cr}(\mathrm{Black} \mid \mathrm{U}_i)]} \\
&= \frac{\left(\frac{1}{4}\right)\left(1\right)}{\left(\frac{1}{4}\right)\left(0\right) + \left(\frac{1}{4}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{4}\right)\left(\frac{3}{4}\right) + \left(\frac{1}{4}\right)\left(1\right)} \\
&= \frac{\frac{4}{16}}{\frac{9}{16}} = \frac{4}{9} \\
&\approx .44
\end{aligned}$$

Similar calculations would show that $\mathrm{Cr}_{\mathrm{new}}(\mathrm{Heads}) = 4/9$, $\mathrm{Cr}_{\mathrm{new}}(\mathrm{U}_1) = 0$, $\mathrm{Cr}_{\mathrm{new}}(\mathrm{U}_2) = 2/9$, and $\mathrm{Cr}_{\mathrm{new}}(\mathrm{U}_3) = 3/9$. Note that given Black, Tails

($= U_2 \vee U_3$) and U_4 are mutually exclusive and jointly exhaustive, and so your credences in them rightly sum to 1.

However, according to IBE*, given that U_4 is the best urn-explanation of your draw, and setting b at 1/8,[23]

$$\mathrm{Cr}_{\mathrm{new}}(U_4) = \frac{\mathrm{Cr}(U_4)\mathrm{Cr}(\text{ Black} \mid U_4) + f(U_4, \text{Black})}{\sum_i [\mathrm{Cr}(U_i)\mathrm{Cr}(\text{Black} \mid U_i) + f(U_i, \text{Black})]} = \frac{\frac{4}{16} + \frac{2}{16}}{\frac{9}{16} + \frac{2}{16}} = \frac{6}{11} \approx .55$$

Similar calculations show that $\mathrm{Cr}_{\mathrm{new}}(U_1) = 0$, $\mathrm{Cr}_{\mathrm{new}}(U_2) = 2/11$, and $\mathrm{Cr}_{\mathrm{new}}(U_3) = 3/11$.

The problem is now easy to see. You know that Heads is true iff U_4 is. However, you assign them different credences: you are more confident than not that the fourth urn was picked but less confident than not that heads was flipped. This is probabilistically incoherent.[24]

So far I have only considered IBE* as a non-Bayesian version of IBE for degrees of belief. In his most recent essay on IBE,[25] Douven considers a more general schema of non-Bayesian forms of IBE of which IBE* is a special case:

$$\mathrm{Cr}_{\mathrm{new}}(H_i) = \frac{\mathrm{Cr}(H_i)\mathrm{Cr}(E \mid H_i) + c \times \mathrm{Cr}(H_i)\mathrm{Cr}(E \mid H_i)\, m(H_i, E)}{\sum_j \left[\mathrm{Cr}(H_j)\mathrm{Cr}(E \mid H_j) + c \times \mathrm{Cr}(H_j)\mathrm{Cr}(E \mid H_j)\, m(H_j, E)\right]}$$

Here $c \in [0, 1]$ is a constant and $m \in [-1, 1]$ is a formal measure of how well H_i explains E. The main difference between this schema and the more specific IBE* is that this schema assigns explanatory bonuses and penalties to all hypotheses, whereas IBE* only assigns a bonus to the

[23] Any non-zero bonus will result in incoherence for the same reason. I choose 1/8 for ease of computation.

[24] IBE* will also lead to incoherence in the Mr. Boddy case. If Plum, Defense, and Planned all receive an explanatory bonus, then $\mathrm{Cr}_{\mathrm{new}}(\text{Plum}) > 2/3$, $\mathrm{Cr}_{\mathrm{new}}(\text{Defense}) > 2/3$, and $\mathrm{Cr}_{\mathrm{new}}(\text{Planned}) > 2/3$. Suppose the jury's new credences are coherent. Then:

$\mathrm{Cr}_{\mathrm{new}}(\text{Plum\&Money\&Planned}) + \mathrm{Cr}_{\mathrm{new}}(\text{Plum\&Defense\&}\sim\text{Planned}) = \mathrm{Cr}_{\mathrm{new}}(\text{Plum}) > 2/3$,

$\mathrm{Cr}_{\mathrm{new}}(\text{Plum\&Defense\&}\sim\text{Planned}) + \mathrm{Cr}_{\mathrm{new}}(\text{Green\&Defense\&Planned}) = \mathrm{Cr}_{\mathrm{new}}(\text{Defense}) > 2/3$,

and

$\mathrm{Cr}_{\mathrm{new}}(\text{Plum\&Money\&Planned}) + \mathrm{Cr}_{\mathrm{new}}(\text{Green\&Defense\&Planned}) = \mathrm{Cr}_{\mathrm{new}}(\text{Planned}) > 2/3$.

But this implies that

$\mathrm{Cr}_{\mathrm{new}}(\text{Plum\&Money\&Planned}) + \mathrm{Cr}_{\mathrm{new}}(\text{Plum\&Defense\&}\sim\text{Planned}) + \mathrm{Cr}_{\mathrm{new}}(\text{Green\&Defense\&Planned}) > 1$,

which is incoherent. So the jury's new credences cannot be coherent.

[25] Douven, "Inference to the Best Explanation: What Is It? And Why Should We Care?," *op. cit.*

most explanatory hypothesis. The constant c determines how much weight is given to explanatory considerations compared to the weight carried by the priors and likelihoods; a higher value gives explanatory considerations greater weight.

In our urn case, this general schema will lead to incoherent results given that $c > 0$, $m(\text{Heads, Black}) < m(\text{Tails, Black})$, $m(U_2, \text{Black}) < m(U_4, \text{Black})$, and $m(U_3, \text{Black}) < m(U_4, \text{Black})$—that is, given that *some* extra weight is given to explanatory considerations, and given that Tails is counted as a better explanation than Heads and U_4 is counted as a better explanation than U_2 or U_3. To see this, note that $Cr_{new}(\text{Heads}) = Cr_{new}(U_4) = 4/9$ if $c = 0$ (in which case this rule reduces to Bayes's rule). If $c > 0$, then the new credence in Tails will be lower, because Tails gets a lower explanatory bonus than Heads, and so after normalizing (by dividing by the denominator) so that $Cr_{new}(\text{Heads}) + Cr_{new}(\text{Tails}) = 1$, $Cr_{new}(\text{Heads}) < 4/9$. Likewise, the new credence in U_4 will be higher, because U_4 gets a bigger explanatory bonus than U_2 and U_3; hence, $Cr_{new}(U_4) > 4/9$. But this is probabilistically incoherent, because Heads and U_4 are equivalent after drawing black and so must have equal probabilities. Thus, we can only avoid incoherence by setting c equal to 0, which reduces this schema to Bayes's rule.

Setting aside the particular schema above, the general problem that I have identified will remain for any credence update rule that gives "bonuses" to explanatorily better hypotheses. Any ranking of explanatory goodness that satisfies (1) will sometimes rank hypotheses at different levels of explanation in ways that lead to inconsistency if explanatory hypotheses get bonuses beyond those already given by likelihoods. This is because the hypothesis ranked best at one level of explanation will not necessarily be consistent with the hypothesis ranked best at another level of explanation.

III. RESPONSES[26]

I have argued that when we are considering potential explanations of the available evidence at multiple levels, following non-Bayesian versions of IBE can lead to either deductive inconsistency or probabilistic incoherence. In this section, I want to consider four responses on behalf of the defender of these versions of IBE. The first response is to deny my claims about which explanations are better in the cases I have described. The second response is that we should not infer or update beliefs or credences in atomic hypotheses of the kind I have considered here but rather in complete world-states. The third response is that the kind of situation I have described is not widespread enough for my argument to undermine most applications of IBE. The fourth

[26] This section benefited greatly from the comments of an anonymous reviewer.

response is to give up deductive consistency and probabilistic coherence as rational requirements on beliefs and credences.

III.1. Denying (1). The most straightforward objection to my initial urn case is that I was wrong to assume that U_4 is a better explanation of Black than U_2 or U_3 or that Tails is a better explanation of Black than Heads. (Analogous remarks go for the Mr. Boddy case; I will focus on the urn case in what follows.) All I have shown, so the objection goes, is that one of these claims about explanatory goodness must be false. I noted in section I that these claims about explanatory goodness follow from

(1) If H_1 and H_2 are both potential explanations of E, $P(H_1) = P(H_2)$, and $P(E|H_1) > P(E|H_2)$, then H_1 explains E better than H_2.

The defender of this first objection must thus deny (1).

In the course of defending the descriptive adequacy of IBE*, Douven and Schupbach[27] consider formalizations of the notion of explanatory goodness employed in IBE* in terms of several proposed probabilistic measures of *explanatory power*.[28] All of these measures imply that

(2) If H_1 and H_2 are both potential explanations of E, $P(H_1) = P(H_2)$, and $P(E|H_1) > P(E|H_2)$, then H_1 is a more powerful explanation of E than H_2.

In fact, they all imply something stronger than (2), namely that

(3) If H_1 and H_2 are both potential explanations of E and $P(E|H_1) > P(E|H_2)$, then H_1 is a more powerful explanation of E than H_2.[29]

If explanatory goodness = explanatory power, then (1) and (2) are equivalent. Hence, if we formalize explanatory goodness in terms of one of the proposed measures of explanatory power in the literature, then we must accept (1). We must also accept the stronger claim that

(4) If H_1 and H_2 are both potential explanations of E and $P(E|H_1) > P(E|H_2)$, then H_1 explains E better than H_2.

[27] Douven and Schupbach, "Probabilistic Alternatives to Bayesianism," *op. cit.*, pp. 3–4.

[28] See, for example, Jonah N. Schupbach, "Comparing Probabilistic Measures of Explanatory Power," *Philosophy of Science*, LXXVIII, 5 (December 2011): 813–29; Jonah N. Schupbach and Jan Sprenger, "The Logic of Explanatory Power," *Philosophy of Science*, LXXVIII, 1 (January 2011): 105–27; Vincenzo Crupi and Katya Tentori, "A Second Look at the Logic of Explanatory Power (with Two Novel Representation Theorems)," *Philosophy of Science*, LXXIX, 3 (July 2012): 365–85.

[29] In fact, most discussions of these measures take for granted even stronger axioms than (3) as conditions of adequacy for a measure of explanatory power.

But one might hold that there are other explanatory virtues besides explanatory power, such as informativeness and fruitfulness. Therefore, H_1 might be a more powerful explanation of E than H_2 and yet still be a worse overall explanation.[30]

However, even if one thought that (1) might be false in cases where, say, H_2 was more informative or fruitful than H_1, the urn case is not such a case. The different coin hypotheses and the different urn hypotheses are symmetrical: not only do they have equal prior probabilities, they are the same kinds of hypotheses, about the same kinds of objects, saying the same kinds of things. None are *ad hoc*. All appear to be equally informative, fruitful, and so on. There seem to be no substantive differences among them except for the degree to which they predict the evidence. In such a case, it is very difficult to deny that the hypothesis that predicts the evidence more strongly is the better explanation of that evidence. Rejecting the above judgments about which explanations are better is thus not an attractive option.

I should note that, although they do not consider examples like those above, Douven and Wenmackers do attempt to preclude the possibility of cases in which IBE* leads to probabilistic incoherence by making a formal assumption about the explanatory bonus function $f(H, E)$:

> [I]t is safe to assume that, for all E, $f(H, E) = 0$ whenever H is a tautology or a contradiction. If we make the further formal assumption that, for all E, $f(H \vee H^*, E) = f(H, E) + f(H^*, E)$ whenever H and H* are mutually exclusive, then it is easy to prove that updating a probability function via IBE* leads again to a probability function.[31]

The urn example violates this second assumption. Your background knowledge implies that Heads is materially equivalent to $U_1 \vee U_4$. But

$$f(\text{Heads, Black}) = 0 \neq f(U_1, \text{Black}) + f(U_4, \text{Black}) = 0 + 1 = 1.$$

[30] My thanks to an anonymous reviewer for pressing me on this point.

[31] Douven and Wenmackers, "Inference to the Best Explanation versus Bayes's Rule in a Social Setting," *op. cit.*, section 2. As stated, these two assumptions are inconsistent. By the first assumption, $f(H \vee \sim H, E) = 0$. But by the second assumption, $f(H \vee \sim H, E) = f(H, E) + f(\sim H, E)$. But if $f(H, E)$ is non-zero, then these two equalities are inconsistent. In correspondence, Douven has suggested that he and Wenmackers should have added the condition that H and H* are not jointly exhaustive. Alternatively, they could limit the first assumption to contradictions alone (because applying IBE* will make $Cr_{new}(H)$ equal to 1 for any value of f, if $Cr(H) = 1$). Both of these fixes have odd consequences—the latter assigns an "explanatory bonus" (albeit a vacuous one) to tautologies, and the former makes f discontinuous in the limit. But because they make no practical difference to credence assignments, neither of these problems seem as serious to me as the inconsistency of the second assumption with our explanatory judgments in the urn case.

Because, as I have argued, it is clear that U_4 is the best explanation among the urn hypotheses and Tails is the best explanation among the coin hypotheses, it follows that the second assumption above is not a tenable constraint on *f* if *f* is a function that assigns a bonus to the best explanation of the evidence.

III.2. World-States. The above inconsistency arose because we were considering two different partitions of hypotheses, at different levels of explanation. If we could ensure that we only consider *one* partition, then there would be no room for a conflict to arise. The easiest way to do this would be to have an agent directly update her beliefs/credences over the partition of complete world-states by IBE, with her new beliefs/credences over world-states imposing beliefs/credences over other partitions. Here a *world-state* is a conjunction such that, for every atomic partition of possibilities in an agent's language explanatorily prior to the observed evidence, the world-state contains one member of each partition as a conjunct. For example, in the original urn scenario, if the atomic partitions prior to the observed evidence Black are {Heads, Tails} and $\{U_1, U_2, U_3, U_4\}$, then there are four world-states consistent with your background knowledge:

$\{\text{Heads}\&U_1, \text{Heads}\&U_4, \text{Tails}\&U_2, \text{Tails}\&U_3\}$.

It follows from (4) that $\text{Heads}\&U_4$ is the best explanation of Black among these hypotheses. If you are fully inferring to the best explanation, then the current proposal would have you infer that $\text{Heads}\&U_4$ is true. You will then believe each of these conjuncts individually as well. If you are updating your credences, then $\text{Heads}\&U_4$ will receive the explanatory boost among the members of this partition. Your other new credences could then be determined by your new credence distribution over this partition, as in a standard probability function: any sentence in an agent's language is equivalent to a disjunction of world-states, and so its probability can be determined by adding the probabilities assigned to those world-states.

I have three objections to this strategy for resolving the problem.[32] The first is that it is impossible to apply in actual reasoning. Outside of idealized thought experiments such as those under discussion here, an agent could not even formulate a world-state, let alone assign a

[32] In section 2 of my manuscript, "The Structure of Epistemic Probabilities" (unpublished), I discuss some additional problems with taking world-states to be the primary objects of inference. (My definition of world-states there differs from that in this paper in that I do not require that the partitions be explanatorily prior to the evidence, but most of the same points apply.)

probability to it. In real life, we rarely consider partitions of complex conjunctions of multiple hypotheses; we instead consider partitions of different atomic or near-atomic propositions.

A second problem for this response is that a maximally specific description of the world is not the kind of thing that we tend to consider a good explanation. Usually when we ask what explains some observed evidence, we are interested in much more specific hypotheses and not in a complete history of the universe. Applying IBE only to world-states is thus in tension with the spirit of IBE, which is usually understood to license inference to these much more specific explanations.

Applying IBE only to world-states is in tension with the spirit of IBE for another reason. The core intuition behind IBE is that how well hypotheses explain the evidence is important for determining whether those hypotheses are true. However, inferring/boosting our credence in the world-state that best explains one's evidence privileges the explanatory relations between world-states and the evidence at the expense of the explanatory relations between atomic propositions and the evidence and the explanatory relations among the atomic propositions themselves. If the fact that Heads&U_4 is the best explanation of Black among {Heads&U_1, Heads&U_4, Tails&U_2, Tails&U_3} gives us reason to infer/boost our credence in Heads&U_4, then it seems that the fact that Tails is the best explanation of Black among {Heads, Tails} should also give us reason to infer/boost our credence in Tails. But non-Bayesian versions of IBE cannot accept both of these claims. It would be preferable to have a version of IBE that can take into account the explanatory goodness of *all* hypotheses, not just that of world-states.

III.3. Limiting IBE. In his most recent essay on IBE, Douven writes that IBE is best thought of as a "slogan," the correct spelling out of which depends on the situation. Another response that the defender of IBE* and IBE for full belief could make would be to grant their inapplicability in the kind of case that I have discussed but to hold that they can still be applied in other cases—that in other cases, they are the best way to spell out the slogan of IBE.[33] A fairly limited restriction would be to hold that these forms of IBE can be applied when no hypotheses at different levels of explanation are inconsistent with each other, as Tails is inconsistent with U_4.

To see whether this restriction works, let us consider a revision of the original urn case. Now, if I flip heads the first time, instead of there

[33] For example, one might support this claim about IBE* with the alleged advantages of IBE* discussed in the texts referenced at the beginning of section II.

being a probability of 0 of selecting U_2 or U_3, I will initiate a chance process that has a 1/100 probability of selecting U_2 and a 1/100 probability of selecting U_3, and that distributes the remaining probability equally among U_1 and U_4. Similarly, if I flip tails, I will initiate a process that has a 1/100 probability of selecting U_1 and a 1/100 probability of selecting U_4, and that distributes the remaining probability equally among U_2 and U_3.

Let us first consider how problematic it is to follow IBE for full belief in this case. It still follows from (1) that Tails is the best coin explanation of drawing black, and U_4 is the best urn-explanation. Because Tails and U_4 are now logically consistent, inferring both of them does not give one deductively inconsistent beliefs. Nevertheless, Tails and U_4 are extremely negatively relevant to each other. Even though neither is very initially unlikely on its own, the prior chance that both are true is only 1/200. That both are true is thus a surprising claim to be committed to by one's inferring to the best explanation of drawing black.

Moreover, the *reason* that Tails is the best coin explanation of Black is in significant tension with U_4. Tails is the best explanation of Black *because* it makes it more likely that $U_2 \vee U_3$, which makes Black more likely than $U_1 \vee U_4$. It would be very odd to infer that U_4 because it explains Black well and to also infer that Tails is true because it explains Black well *by making likely hypotheses that one is rejecting in inferring U_4*. As with the world-state revision considered above, inferring both of these would ignore the negative explanatory relation that Tails and U_4 bear toward each other in favor of the positive explanatory relations that they individually bear toward Black.

Finally, inferring the best explanation in this but not in the original urn case would commit us to a problematic discontinuity as we move from extreme negative relevance to logical inconsistency. Inconsistency is a limiting case of negative relevance, and it seems bad to allow the inference of two propositions as they become more and more negatively relevant to each other but to disallow it as soon as they reach the point of inconsistency.

As for IBE*, it still leads to incoherence. In this case, we have the following hypotheses and (initial) credences:

U_1 = the urn selected contains 4 white balls
U_2 = the urn selected contains 2 black balls, 2 white balls
U_3 = the urn selected contains 3 black balls, 1 white ball
U_4 = the urn selected contains 4 black balls
$Cr(U_1|\text{Heads}) = Cr(U_4|\text{Heads}) = 49/100$
$Cr(U_2|\text{Heads}) = Cr(U_3|\text{Heads}) = 1/100$
$Cr(U_1|\text{Tails}) = Cr(U_4|\text{Tails}) = 1/100$

$Cr(U_2|Tails) = Cr(U_3|Tails) = 49/100$

$Cr(Black|Heads) = Cr(U_1|Heads)Cr(Black|U_1) + Cr(U_2|Heads)Cr(Black|U_2) + Cr(U_3|Heads)Cr(Black|U_3) + Cr(U_4|Heads)Cr(Black|U_4) = (49/100)(0) + (1/100)(1/2) + (1/100)(3/4) + (49/100)(1) = 1/200 + 3/400 + 49/100 = 201/400 = .5025$

$Cr(Black|Tails) = Cr(U_1|Tails)Cr(Black|U_1) + Cr(U_2|Tails)Cr(Black|U_2) + Cr(U_3|Tails)Cr(Black|U_3) + Cr(U_4|Tails)Cr(Black|U_4) = (1/100)(0) + (49/100)(1/2) + (49/100)(3/4) + (1/100)(1) = 49/200 + 147/400 + 1/100 = 249/400 = .6225$

As noted above, it still follows from (1) that Tails is the best coin-explanation of drawing black, and U_4 is the best urn-explanation. If we consider the set of world-states {Heads&U_1, Heads&U_2, Heads&U_3, Heads&U_4, Tails&U_1, Tails&U_2, Tails&U_3, Tails&U_4}, it follows from (4) that Heads&U_4 and Tails&U_4 are better explanations of Black than are any of the other world-states. If we identify explanatory goodness with explanatory power, it will also follow from any of the measures of explanatory power mentioned earlier that Heads&U_4 and Tails&U_4 are equally good explanations, because they make the evidence equally probable. In such a case, Douven and Wenmackers say that Heads&U_4 and Tails&U_4 should split the explanatory bonus b equally, so that $f(H\&U_4,B) = f(T\&U_4,B) = b/2$.[34]

It then follows from IBE* that

$$Cr_{new}(Heads) = \frac{Cr(Heads)Cr(Black \mid Heads)}{Cr(Heads)Cr(Black \mid Heads) + Cr(Tails)Cr(Black \mid Tails) + b} = \frac{\frac{1}{2}\left(\frac{201}{400}\right)}{\frac{1}{2}\left(\frac{201}{400}\right) + \frac{1}{2}\left(\frac{249}{400}\right) + b} = \frac{\frac{201}{800}}{\frac{450}{800} + b} = \frac{\frac{201}{800}}{\frac{9}{16} + b}$$

$$Cr_{new}(Heads\&U_1) + Cr_{new}(Heads\&U_2) + Cr_{new}(Heads\&U_3) + Cr_{new}(Heads\&U_4)$$

$$= \frac{Cr(Heads\&U_1)Cr(Black \mid Heads\&U_1)}{Cr(Black) + b} + \frac{Cr(Heads\&U_2)Cr(Black \mid Heads\&U_2)}{Cr(Black) + b}$$

$$+ \frac{Cr(Heads\&U_3)Cr(Black \mid Heads\&U_3)}{Cr(Black) + b} + \frac{Cr(Heads\&U_4)Cr(Black \mid Heads\&U_4) + \frac{b}{2}}{Cr(Black) + b}$$

$$= \frac{\frac{49}{200}(0) + \frac{1}{200}\left(\frac{1}{2}\right) + \frac{1}{200}\left(\frac{3}{4}\right) + \frac{49}{200}(1) + \frac{b}{2}}{\frac{9}{16} + b} = \frac{\frac{201}{800} + \frac{b}{2}}{\frac{9}{16} + b}$$

[34] Douven and Wenmackers, "Inference to the Best Explanation versus Bayes's Rule in a Social Setting," *op. cit.*, section 2.

If your new credence distribution is coherent, then $Cr_{new}(\text{Heads}) = Cr_{new}(\text{Heads\&}U_1) + Cr_{new}(\text{Heads\&}U_2) + Cr_{new}(\text{Heads\&}U_3) + Cr_{new}(\text{Heads\&}U_4)$. However,

$$\frac{\frac{201}{800}}{\frac{9}{16}+b} = \frac{\frac{201}{800}+\frac{b}{2}}{\frac{9}{16}+b}$$

iff $b = 0$. As such, your new credence distribution is coherent iff $b = 0$, in which case IBE* reduces to Bayes's rule. Therefore, applying IBE* with a non-zero explanatory bonus in this new case still leads to incoherence.

Douven's more general schema will also lead to incoherence. It will give an explanatory penalty to Heads, thus lowering $Cr_{new}(\text{Heads})$. However, it will give equal penalties/bonuses to Heads&U_2 and Tails&U_2, Heads&U_3 and Tails&U_3, and Heads&U_4 and Tails&U_4, respectively. ($Cr_{new}(\text{Heads\&}U_1) = Cr_{new}(\text{Tails\&}U_1) = 0$ no matter what.) This will increase the sum

$$Cr_{new}(\text{Heads\&}U_1) + Cr_{new}(\text{Heads\&}U_2) + Cr_{new}(\text{Heads\&}U_3) + Cr_{new}(\text{Heads\&}U_4).$$

This is because, *without* the bonus b, this sum is less than the sum

$$Cr_{new}(\text{Tails\&}U_1) + Cr_{new}(\text{Tails\&}U_2) + Cr_{new}(\text{Tails\&}U_3) + Cr_{new}(\text{Tails\&}U_4),$$

and adding the same quantity to two values makes their ratio closer to equal, thus making the normalized sum of the former credences higher. Therefore,

$$Cr_{new}(\text{Heads}) > Cr_{new}(\text{Heads\&}U_1) + Cr_{new}(\text{Heads\&}U_2) + Cr_{new}(\text{Heads\&}U_3) + Cr_{new}(\text{Heads\&}U_4),$$

which is incoherent.

Applying non-Bayesian forms of IBE remains problematic, then, even when two best explanations at different levels of explanation are merely negatively relevant to each other and not inconsistent. So, if we are to apply these forms of IBE only to unproblematic cases, then we must limit their application to cases in which we are either not interested in multiple levels of explanation or in which the best explanations at different levels are not negatively relevant to each other.

Although I do not think that my argument in this paper definitively rules out the viability of such a restriction, I do want to suggest two reasons to be wary of it. First, it would be preferable, all other things equal, to have a unified inference form that we can use in all or most contexts. If we think that this inference form will be a version of IBE,

then we should prefer a Bayesian explication of IBE (such as the one gestured at in section IV) that does not lead to inconsistency or incoherence in any cases.

Second, the above restricted forms of IBE could arguably only rarely be applied, inasmuch as the phenomenon of reasoning about multiple levels of explanation is quite common. Although the urn case is clearly a toy example, the Mr. Boddy example is fairly realistic (aside from the board game references). The jury on a normal homicide case really *will* be interested not only in the question of who the killer is but also in the circumstances surrounding the killing, such as the killer's motives, which is a level of explanation up from the question of who the killer is. This information is important because a revenge killing, but not a self-defense killing, would constitute murder. In addition, if the killing does constitute murder, then the jury needs to know how severe a sentence the circumstances warrant, and so they need to determine whether the murder was premeditated.

As another example, imagine a doctor observing particular symptoms, such as headaches and chest pain, in a patient. He considers certain physiological conditions that might cause these symptoms, such as high blood pressure and bronchitis. He further considers certain behavioral patterns that are likely to have caused these conditions, such as poor diet and smoking cigarettes. It is important that he determine the correct explanation at *both* of these levels so as to best judge how to treat the condition that the patient is most likely to have.

Reasoning about multiple levels of explanation, then, is a familiar feature of ordinary life.[35] One might still think that only rarely will the best explanations at different levels be negatively relevant to each other, so that we can still apply the above forms of IBE in most cases of reasoning about multiple levels of explanation. This claim is difficult to assess from the armchair. But even if in many cases the best

[35] The kind of complexity present in multiple levels of explanation is really a specific form of the more general Quinean phenomenon in which all of our beliefs are connected. Formal epistemologists have largely neglected this phenomenon, perhaps because of the difficulty of formally representing it. In my own view, the best formal representation of this interconnectedness is found in the theory of Bayesian networks, as developed by Judea Pearl in his *Probabilistic Reasoning in Intelligent Systems* (San Mateo, CA: Morgan Kaufmann, 1988). Bayesian networks have been applied to epistemology in Luc Bovens and Stephan Hartmann's *Bayesian Epistemology* (Oxford: Oxford University Press, 2003); Leah Henderson et al.'s "The Structure and Dynamics of Scientific Theories: A Hierarchical Bayesian Perspective," *Philosophy of Science*, LXXVII, 2 (April 2010): 172–200; and in my "The Structure of Epistemic Probabilities," *op. cit.* These texts all contain further examples of empirical and scientific reasoning that is concerned with more than one level of explanation.

explanations at multiple levels will not be negatively relevant to each other, we will often not be able to know this fact in advance of investigation. For example, we can imagine that as the doctor learns more, he comes to conclude that the best physiological explanation of the patient's headache is high blood pressure, and the best behavioral explanation of high blood pressure is smoking, but smoking is not the best behavioral explanation of the patient's headache. Inasmuch as we usually cannot rule out such a situation in advance, we should be wary of using a rule that will run into trouble in such a situation.

III.4. Giving Up Consistency/Coherence. I have argued that following non-Bayesian forms of IBE sometimes leads either to deductive inconsistency, in the case of IBE for full belief, or probabilistic incoherence, in the case of non-Bayesian versions of IBE for degrees of belief. However, I have not yet said much about why this is a negative consequence.

Consider probabilistic incoherence first. Philosophers disagree about whether rational credences must be diachronically coherent. However, most philosophers who work on the subject agree that rational credences must be *synchronically* coherent. The reasons for this belief perhaps differ from philosopher to philosopher (see section IV for one basis for synchronic coherence that does not extend to diachronic coherence). This is not the place to rehearse the various justifications commonly given for probabilistic coherence. It is enough to note that IBE* has so far been explored in contexts that take for granted the need for synchronic coherence. The papers by Douven, Schupbach, and Wenmackers cited at the beginning of section II all assume that credences updated by IBE* will be probabilities, and their discussions tend to take for granted that this is a condition of an adequate update rule. For example, Douven and Wenmackers write that "we will want an update rule to be formally adequate, at least in that it outputs a probability function when given a probability function as input."[36] Thus, the argument of this paper shows that IBE* does not satisfy one of the criteria of adequacy assumed in most discussions of what form a credal version of IBE should take. Dialectically, then, it shows that IBE* is not the right way to spell out the slogan of IBE.

Consider deductive consistency next. It is more common for philosophers to give up this requirement in light of apparent counterexamples. Many philosophers have argued that because of the preface and lottery paradoxes, anyone who thinks that it is sometimes rational to believe propositions that are less than fully certain must accept that

[36] Douven and Wenmackers, "Inference to the Best Explanation versus Bayes's Rule in a Social Setting," *op. cit.*, section 2.

it is sometimes rational to believe propositions that are jointly inconsistent.[37] In light of this argument, one might take my argument to simply give us another reason to give up deductive consistency as a requirement for rational belief.

In response, I would like to make two points. First, my argument may still be effective for philosophers who believe that the preface and lottery paradoxes can be resolved in such a way as to not give up on deductive consistency as a requirement for full beliefs. These philosophers should agree that the fact that following the above versions of IBE sometimes leads to inconsistent beliefs gives us reason to reject these versions of IBE.[38]

Second, even if we can rationally have a large inconsistent set of beliefs, it remains plausible that it is never rational to believe *two* propositions that are inconsistent with each other. For example, Easwaran and Fitelson write that "smaller inconsistent belief sets seem 'less coherent' than larger inconsistent belief sets."[39] Only for inconsistent sets of size 2 do they think it completely clear that rationality precludes belief in all the members of the set. Therefore, if we are to retain IBE for full belief by giving up on consistency, we should arguably adopt a version of IBE on which, in order for an explanation to count as good enough to infer, it must have a probability of at least .5.

[37] See, for example, David Christensen, *Putting Logic in its Place* (Oxford: Oxford University Press, 2004); Kenny Easwaran and Branden Fitelson, "Accuracy, Coherence, and Evidence," in Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, vol. 5 (New York: Oxford University Press, 2015), pp. 61–96.

[38] One might reply that we could borrow from these philosophers' resolutions of the preface and lottery paradoxes to add a requirement to IBE for full belief that would rule out the kinds of inferences I discussed. For instance, if these philosophers think you should not hold lottery beliefs because they have property X, and the IBE-based beliefs in my cases have property X, then we can add to IBE the requirement that one not infer a conclusion that has property X. Although I cannot consider all of the resolutions of the lottery and preface paradoxes that have been proposed here, I will note that at least some will not extend to my cases. For example, Dana Nelkin tries to maintain deductive consistency in the face of the lottery paradox by arguing that rational belief cannot be based on purely statistical evidence, as one's belief that one will lose a fair lottery is if one bases it purely on the number of tickets in the lottery ("The Lottery Paradox, Knowledge, and Rationality," *Philosophical Review*, CIX, 3 (July 2000): 373–409). Rather, rational belief requires that one see an apparent "causal or explanatory connection between [one's] belief and the fact that makes it true" (*ibid.*, p. 396). However, beliefs based on inference to the best explanation are the paradigm cases in which there is an apparent causal or explanatory connection between one's belief and the fact that makes it true. For instance, if you infer that U_4 because you drew a black ball out of the urn and U_4 posits the greatest number of black balls in the urn, then your belief is based on evidence that U_4, if true, helped make true. Likewise, if the jury infers that Plum is guilty because Boddy was killed with the revolver and Plum was more likely to use the revolver than Green was, their belief is based on evidence that Plum's being the murderer made true, if Plum is the murderer.

[39] Easwaran and Fitelson, "Accuracy, Coherence, and Evidence," *op. cit.*, pp. 83–84.

This narrows down the versions of IBE to which my argument applies, but it is still an interesting result that we should reject versions of IBE for full belief that do not include or imply this threshold requirement. In addition, inasmuch as the argument of section II shows that the relevant probabilities for this requirement must be determined in a traditional Bayesian way[40] rather than according to IBE* or some other non-Bayesian updating rule, this form of IBE will be, in some sense, a Bayesian version of IBE. My argument will then work against versions of IBE that are not Bayesian in this sense.

IV. CONCLUSION

In this paper, I have argued that non-Bayesian forms of IBE sometimes lead either to deductive inconsistency, in the case of IBE for full belief, or probabilistic incoherence, in the case of versions of IBE for degrees of belief that use non-Bayesian updating rules. They do so in situations where we are concerned with multiple levels of explanation, and the best explanations at different levels conflict with each other. Inasmuch as this phenomenon is both common and problematic, we should reject these forms of IBE as general rules for inference or updating.

In spite of these arguments, I endorse the claim that explanatory considerations are central to good inference. Having argued against non-Bayesian forms of IBE, in this last section, I want to briefly consider how explanatory factors could inform inference within a Bayesian framework.

First, note that even if we conclude from the above that we must reject any form of IBE that is inconsistent with Bayesian conditionalization, explanatory considerations could still be relevant to confirmation inasmuch as, frequently,

$$P(H|E\&[\text{there is an explanatory connection between E and H}]) > P(H|E\&[\text{there is no explanatory connection between E and H}]).$$[41]

But in fact I do *not* think the above arguments show that we should reject any form of IBE that is inconsistent with Bayesian conditionalization. There are already good reasons to reject Bayes's rule as a universal updating rule. As philosophers have long argued,[42] there

[40] But see section IV, where I suggest that explanatory factors do play a role in determining these probabilities, just not the kind of role embodied in a non-Bayesian updating rule.

[41] I argue for this claim in my "How Explanation Guides Confirmation," *Philosophy of Science*, LXXXIV, 2 (April 2017): 359–68, against William Roche and Elliot Sober, "Explanatoriness Is Evidentially Irrelevant, or Inference to the Best Explanation Meets Bayesian Confirmation Theory," *Analysis*, LXXIII, 4 (October 2013): 659–68.

[42] For example, Fahiem Bacchus, Henry E. Kyburg Jr., and Mariam Thalos, "Against Conditionalization," *Synthese*, LXXXV, 3 (December 1990): 475–506.

are many cases in which it is clear that we should not update by conditionalization, such as when we start out with (what we now recognize to be) irrational credences.

I endorse the traditional Bayesian objection to formulations of IBE for full belief that they do not take into account degrees of confidence or degrees of confirmation. However, given the above problems with conditionalization, my objection to IBE* is not the obvious Bayesian one that it violates conditionalization. Rather, I think that in a sense, IBE* does not go far enough. IBE* (and the more general schema considered in section II) replaces a rule that is a complete function of one's prior credence distribution with one that is a partial function of one's prior credence distribution. Standard Bayesian conditionalization "inherits" the coherence of your past credence state: it keeps your credences in propositions at different levels of explanation coherent by preserving the coherent relations between them in your past credence function. By distorting these relations, IBE* leads to probabilistic incoherence. But making your current credences a function of your past credences is not the only way to get probabilistic coherence. Another way is to make them a function of the *objective epistemic probabilities*.

Hence, in rejecting Bayes's rule, I think we should go further than IBE*. We should give up on the idea that one's current credences are in any way beholden to one's past credences, and so give up on rules for *updating* altogether. If you start out with irrational credences, no update rule will save you. You should base your current credence in H, not on the idiosyncratic opinions of your past self about H, but on the objective epistemic probability of H given your evidence. And, as Jonathan Weisberg argues,[43] it is *here* that explanatory factors can play an important role—in determining the correct *a priori* probability distribution. The correct explanationist alternative to Bayes's rule is not a non-Bayesian updating or inference rule but, rather, an objective form of Bayesianism in which explanation plays a central role. Exploring that role is an important task for future research on IBE and Bayesianism.

NEVIN CLIMENHAGA

University of Notre Dame

[43] Jonathan Weisberg, "Locating IBE in the Bayesian Framework," *Synthese*, CLXVII, 1 (March 2009): 125–43.